Journal of Applied Research in Memory and Cognition 8 (2019) 118–129

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

When the Unlikely Becomes Likely: Qualifying Language Does Not Influence Later Truth Judgments

Matthew L. Stanley*, Brenda W. Yang, and Elizabeth J. Marsh

Duke University, United States

Judgments and decisions are frequently made under uncertainty. People often express and interpret this uncertainty with epistemic qualifiers (e.g., *likely*, *improbable*). We investigate the extent to which qualifiers influence truth judgments over time. In four studies, participants studied qualified statements, and two days later they rated the truth of previously qualified statements along with new statements. Previously qualified statements were rated as more likely true than new statements, even when the qualifiers had distinctly opposite meanings (i.e., *certain* versus *impossible*; Study 1) and when all qualifiers cast doubt on the veracity of the statements (e.g., *improbable*, *impossible*; Studies 2–4). Three additional studies suggested that this effect was not dependent on memory for the qualifiers. Consistent with a fluency interpretation, prior exposure made the statements easier to read, driving truth judgments, and overriding the influence of qualifying information. Implications for improving communication using qualifiers are discussed.

General Audience Summary

In our daily and professional lives, we frequently make judgments and decisions under uncertainty. We often use words such as *likely*, *improbable*, and *possible* to express uncertainty; these words are commonly called "qualifying terms." We know how these qualifying terms are typically interpreted when people read them in the moment, but we do not currently know the extent to which qualifiers influence how much people believe statements over time. In our studies, participants studied qualified statements, and two days later they rated the truth of those statements as well as new ones. Previously qualified statements were rated as more likely to be true than new statements, even when the original qualifiers had clearly opposite meanings (i.e., *certain* versus *impossible*) and when all original qualifiers cast doubt on the veracity of the statements (i.e., *uncertain*, *unlikely*, *improbable*, or *impossible*). Previously seen statements were rated as equally likely to be true, no matter how the statements were previously qualified. To the extent qualifying information no longer informs truth judgements after only two days, these results have implications for people's ability to make informed, accurate judgments and decisions in many situations. Our findings have particularly serious ramifications for decision makers in high-stakes contexts (e.g., those working in medicine, law, finance, and intelligence), where it is common for stakeholders to rely on qualifiers when sharing and interpreting critical information.

Keywords: Truth, Knowledge, Memory, Linguistic, Epistemic

Author Note

Brenda W. Yang, Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Duke University, United States; Elizabeth J. Marsh, Department of Psychology and Neuroscience, Duke University, United States.

This research was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation.

* Correspondence concerning this article should be addressed to Matthew L. Stanley, Department of Psychology & Neuroscience, Duke University, 417 Chapel Drive, Durham, NC 27708, United States. Contact: mls103@duke.edu

In many real-world situations, we must make judgments and decisions in the face of uncertainty. Some clinical diagnoses, for example, are based on symptom clusters rather than a definitive test (e.g., ADHD or fibromyalgia). The physician's responsibility is to clearly and sensibly convey the likelihood that a patient has a certain disease or disorder. Moreover, physicians' assessments of treatment efficacy are also often communicated to patients with some degree of uncertainty (e.g., *It's possible that this new medication will alleviate the swelling*). In such situations, physicians and patients rely on *epistemic qualifiers* such as *probable*, *possible*, *unlikely*, and *certain* to express and comprehend different likelihoods that information is true (Brun & Teigen, 1988; Caton, 1966; Wallsten, 1990). Another example comes from the intelligence community, where specific guidelines were developed to map qualifying language onto numerical odds (Heuer, 1999). These guidelines were meant to help officials appropriately respond to threats and crises (e.g., *It's probable that the insurgents are mobilizing* might refer to a 75% chance, give or take 12%). Medicine and intelligence are but two of many fields in which epistemic qualifiers are used extensively, and this qualifying language allows for a relatively easy, natural mode of communication (Wallsten, Budescu, Zwick, & Kemp, 1993). Considerable past research has characterized the psychological principles underlying the representation, use, and communication of epistemic qualifiers in various contexts (e.g., Budescu & Wallsten, 1995; Dhami & Wallsten, 2005; Teigen & Brun, 2000; Wallsten, Budescu, & Tsao, 1997). In contrast, our focus is on the longer-term consequences of qualifying a statement. That is, do later judgments of the veracity of claims reflect how those claims were originally qualified?

To make predictions about the long-term effects of qualifying language, it is useful to consider more generally how people evaluate the truth of information. People often rely on heuristics when judging truth, interpreting easy processing or fluency as evidence for truth (Boehm, 1994; Unkelbach, 2007). As a result, easy-to-read statements (i.e., in high-contrast font) are judged as more likely to be true than ones in a hard-to-read low-contrast font (Parks & Toth, 2006; Reber & Schwarz, 1999), and rhyming statements are judged as more likely to be true than non-rhyming ones (McGlone & Tofighbakhsh, 2000). Moreover, dozens of studies have shown that people are more likely to believe that a previously seen or heard statement is true compared with a novel statement (*illusory truth effect*; Dechêne, Stahl, Hansen, & Wänke, 2010; Hasher, Goldstein, & Toppino, 1977). The effects of repetition are robust, increasing the perceived truth of information across diverse knowledge domains and for long delays (for a review, see Dechêne et al., 2010).

Notably, illusory truth effects do not require verbatim repetitions of statements. That is, prior exposure to part of a claim (e.g., the phrase "a hen's body temperature") is enough to increase belief in a more specific claim (e.g., "the temperature of a hen's body is about 104° F"; Begg, Armour, & Kerr, 1985; see also Arkes, Boehm, & Xu, 1991). This result is consistent with the larger priming literature, where exact repetition of words and phrases is not required to speed processing of related content. For instance, a word is processed more quickly following exposure to a different form of a verb (e.g., *jumped* primes *jump* and *found* primes *find*; Marslen-Wilson & Tyler, 1997) or different but related words (e.g., *bread* primes *butter*; Meyer & Schvaneveldt, 1971). Exposure to two content words yields almost as much priming as does reading the entire sentence containing those words (e.g., *skier*, *buried* yields almost as much priming of *snow* as does *The skier was buried in the snow*; West & Stanovich, 1988). Thus, a qualified version of a statement (e.g., *It's unlikely that the body of a rotten tree is called a daddock*) has the potential to create a fluency signal by speeding later processing of an unqualified version of the statement (e.g., *the body of a rotten tree is called a daddock*).

Of course, fluency is not always used as a cue for truth, but the conditions under which people discount fluency as a guide to truth are less clear. On the one hand, people are less likely to rely on fluency when the source of the repeated statements lacks credibility (Begg, Anas, & Farinacci, 1992; Brown & Nix, 1996). However, fluency affects truth judgments even when relevant knowledge is stored in memory: reading falsehoods such as "*A sari is the short pleated skirt worn by Scotsmen*" increases later truth ratings (Fazio, Brashier, Payne, & Marsh, 2015). And in other situations, people rely on both fluency and other contextual information to make truth judgments. In one study examining biasing comments, for example, both positively (e.g., "It is frequently said that...") and negatively biased statements (e.g., "Few would believe...") were later rated as more likely to be true than new statements after a delay of a few minutes, but positively biased statements were rated as more likely to be true than those that had been negatively biased (Begg & Armour, 1991).

We build on this prior work with an examination of the longer-term consequences of qualifying statements, as is more common in the real world where people regularly retrieve and use previously learned information days, weeks, or even months later. In the studies that follow, participants read qualified statements. They then judged the truth of those statements without the qualifying information, as well as new statements. Study 1 used both positive (e.g., *certain*, *probable*) and negative (e.g., *unlikely*, *impossible*) epistemic qualifiers, while the subsequent studies used only negative (e.g., *improbable*, *impossible*) epistemic qualifiers.

There are reasons to suspect that epistemic qualifiers will have a significant influence on later judgments of truth above and beyond the effect of simply repeating statements (i.e., the illusory truth effect). Epistemic qualifiers are frequently employed and relied upon in everyday and professional life (Budescu & Wallsten, 1995; Wallsten et al., 1997), and intra-individual interpretations of qualifiers are relatively consistent and reliable, especially for anchoring terms like *certain* and *impossible* (Beth-Marom, 1982; Brun & Teigen, 1988; Budescu & Wallsten, 1995; Dhami & Wallsten, 2005; Reagan, Mosteller, & Youtz, 1989). On the other hand, memory for the qualifiers likely becomes worse over time, potentially leaving fluency as the driver of truth judgements (as occurs when people forget that information comes from a low credibility source; Mitchell, Dodson, & Schacter, 2005). To directly assess this possibility, we implemented statistical equivalence testing to determine whether later truth judgments were effectively equivalent, regardless of the qualifier originally accompanying the statement. Furthermore, we directly tested the role of memory for the qualifiers in several paired studies (Studies 2b, 3b, and 4b). The purpose of these three additional studies is to determine whether the qualifying information is more likely to influence later truth judgments when participants better remember the qualifying information.

Study 1

In Study 1, we examined how both positive (i.e., *certain*, *probable*) and negative (i.e., *unlikely*, *impossible*) epistemic qualifiers influence later truth judgments.

Method

Participants. 131 individuals voluntarily participated in this study via Amazon's Mechanical Turk (AMT) and completed both sessions (attrition rate = 18%). Participant recruitment was restricted to individuals in the United States with a prior approval rating above 80%. One participant was excluded for failing to follow directions, so the data were analyzed with the remaining 130 individuals ($M_{\text{age}} = 39.19$, $SD = 12.73$, range: 18–74, 66 females). All participants reported being fluent English speakers. The sample size was determined by using a power analysis for equivalence testing of paired samples, as advocated by Chow, Wang, and Shao (2007). The Duke University Campus Institutional Review Board approved all procedures for all studies reported.

Materials. Stimuli consisted of 160 true unfamiliar statements that had been validated and used in previous research (Wang et al., 2016). Statements were qualified during the initial exposure phase with *certain*, *unlikely*, *probable*, or *impossible*. For example, the statement "The body of a rotten tree is called a daddock" was qualified as "It's unlikely that the body of a rotten tree is called a daddock." These were randomized across participants such that each statement was accompanied by each qualifier for some subset of participants.

Procedure. After providing informed consent, participants completed the initial *exposure phase*. Participants rated 80 statements (each of the four qualifiers accompanied 20 different statements) for subjective interest on an 8-point scale ranging from 1 (*very interesting*) to 8 (*very uninteresting*). Response submission for each statement was only available after 3 s had passed to encourage processing. Participants were explicitly told that the qualifiers provided information about the likelihood of each statement being true or false. Participants were told not to use any outside resources to help them with the task. After rating interest for all statements, participants were asked to self-report if they had prior knowledge about at least one of the factual statements used.

Two days after the initial exposure phase, participants completed the *truth rating phase*. Once it became available on AMT, participants were reminded via email to participate in this second session. Participants rated a total of 160 statements (80 old and 80 new) on likelihood of being true, using an 8-point scale ranging from 1 (*definitely false*) to 8 (*definitely true*). None of the statements in the truth rating phase were qualified, regardless of whether they were old or new. For example, if a participant saw "It's likely that the body of a rotten tree is called a daddock" in the exposure phase, they rated "The body of a rotten tree is called a daddock" in the truth rating phase. Response submission for each statement was only available after 3 s had passed. Participants were instructed not to use any outside resources to help them with the task; this behavior was probed at the end of the experiment, with the assurance that answering truthfully would not affect compensation. The one participant who reported having used outside resources was excluded, as described above.

Equivalence testing. Traditional statistical tests of differences (e.g., *t* and *F* tests) are appropriate for assessing possible differences between conditions or groups. A non-significant finding using null hypothesis significance testing (NHST) does not indicate that conditions or groups are the same (Mara & Cribbie, 2012), and one cannot support the hypothesis that a true effect size is exactly zero (Lakens, 2017). In contrast, statistical equivalence tests are used when the objective is to identify two conditions or groups that are nearly equivalent on some outcome, such that any difference in means is small enough to be considered inconsequential. Statistical equivalence testing is not conceptually or mathematically the same as identifying null results in a NHST framework. We employed this method of analysis to investigate the possibility that truth ratings for statements originally presented with epistemic qualifiers would be rated equivalently in the second session.

A commonly used approach is to test for equivalence of means using the two one-sided test (TOST) procedure (Schuirmann, 1987). The goal in the TOST approach is to specify a lower and upper bound, such that results falling within this range are deemed equivalent to the absence of an effect that is worthwhile to examine (Mara & Cribbie, 2012). This interval represents how far two conditions can be apart and still be considered equivalent. The null hypothesis is defined by two simultaneous predictions that both must be rejected in order to declare the mean differences in paired observations equivalent (where "equivalent" is defined in terms of the established equivalence interval). H_{01} would be rejected if $t_1 \leq -t_{\alpha,n-1}$ and H_{02} would be rejected if $t_2 \leq -t_{1-\alpha,n-1}$, where

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\frac{s_{Diff}}{\sqrt{n-1}}} \quad \text{and} \quad t_2 = \frac{\bar{x}_1 - \bar{x}_2 - (-\delta)}{\frac{s_{Diff}}{\sqrt{n-1}}}$$

Here, $\bar{x}_1 - \bar{x}_2$ are the sample means, $(-\delta, \delta)$ is the specified equivalence interval, and s_{Diff} is the standard deviation of the difference scores. A significant finding using equivalence testing indicates the groups are equivalent (Mara & Cribbie, 2012). Thus, equivalence testing for paired samples was implemented to investigate whether truth ratings were equivalent regardless of how statements were originally qualified. Given the range of reported effect sizes from the meta-analysis conducted by Dechêne et al. (2010), the lower bound for equivalence testing was set at Cohen's $d = -.30$ and the upper bound was set at Cohen's $d = .30$. The typical size of the illusory truth effect—with exposure and truth ratings separated by one day

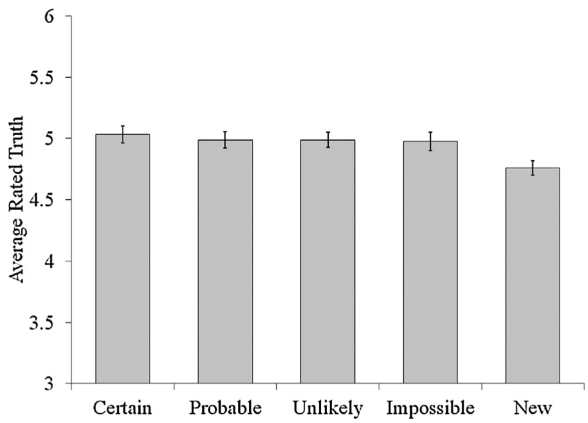

Figure 1. For Study 1, means and standard errors are depicted for the rated truth of both new statements and old statements grouped by the qualifier originally seen accompanying the statement.

to a week—is $d = .43$ (95% CI [.32, .54]). Moreover, an effect size magnitude less than a Cohen's d of .30 is considered rather small by current convention. All statistical equivalence testing was conducted using R with the TOSTER software package (Lakens, 2017).

Results and Discussion

The alpha level for all statistical tests was set at .05. Eighty-eight percent of participants reported not knowing whether any of the statements were true or false in the initial exposure phase. Furthermore, our pattern of results did not differ after excluding participants who reported knowing whether at least one of the statements was true or false.

We first investigated whether old, previously seen statements—regardless of how those statements were originally qualified—were given higher truth ratings than new statements. A paired t test revealed that the average rated truth of old, previously qualified statements ($M = 5.00$, $SD = .72$) was significantly higher than that of new statements ($M = 4.76$, $SD = .68$), $t(129) = 6.97$, $p < .001$, 95% CI [.171, .307].[1] New statements first presented in the truth rating phase were not statistically equivalent in rated truth to old, previously qualified statements ($df = 129$; $t_1 = -10.41$, $p_1 < .001$; $t_2 = -3.57$, $p_2 > .999$).

Next, we investigated whether truth ratings were statistically equivalent regardless of which qualifier accompanied the statements in the exposure phase (see Figure 1). To preview, truth ratings were statistically equivalent regardless of how those statements were originally qualified. Specifically, statements originally qualified with *certain* were statistically equivalent in rated truth to statements that had been qualified with *unlikely* ($df = 129$; $t_1 = -2.42$, $p_1 = .008$; $t_2 = 4.42$, $p_2 < .001$), *probable* ($df = 129$; $t_1 = -2.28$, $p_1 = .012$; $t_2 = 5.96$, $p_2 < .001$), or *impossible* ($df = 129$; $t_1 = -2.29$, $p_1 = .012$; $t_2 = 4.55$, $p_2 < .001$). Similarly, statements previously qualified with *unlikely* were statistically equivalent in rated truth to statements previously qualified with *probable* ($df = 129$; $t_1 = -3.44$, $p_1 < .001$; $t_2 = 3.40$, $p_2 < .001$) or *impossible* ($df = 129$; $t_1 = -3.16$, $p_1 < .001$; $t_2 = 3.68$, $p_2 < .001$). Finally, statements previously qualified with *probable* were statistically equivalent in rated truth to statements previously qualified with *impossible* ($df = 129$; $t_1 = -3.19$, $p_1 < .001$; $t_2 = 3.65$, $p_2 < .001$).

A subsequent ANOVA was computed using a single within-subjects factor (qualifier during the exposure phase: *certain*, *unlikely, probable*, or *impossible*). The dependent variable was the rated truth of statements provided during the truth rating phase. Mauchly's Test of Sphericity indicated that the assumption of sphericity was violated ($\chi^2(5) = 19.33$, $p = .002$), so we implemented a Greenhouse–Geisser correction. There was no significant effect of qualifier on the rated truth of statements, $F(2.70, 348.66) = .66$, $p = .56$, partial $\eta^2 = .005$ (see Figure 1). Subsequent tests of pair-wise comparisons revealed no significant differences in rated truth as a function of the particular qualifier accompanying the statements during the exposure phase before correction for multiple comparisons (all $ps > .25$).

In sum, the rated truth of old statements was statistically equivalent regardless of how they were originally qualified. Prior exposure mattered—old, repeated statements were rated as more likely to be true than new statements—but this illusory truth effect persisted regardless of how the statements were initially qualified. These findings are surprising given that epistemic qualifiers like *certain* and *impossible* hold precise and opposite meanings: what is *certain* should be true, and what is *impossible* should be false. In other words, while some linguistic, epistemic qualifiers are vague, interpretations of anchoring terms like *certain* and *impossible* tend to be highly consistent and reliable within and between individuals (Beth-Marom, 1982; Brun & Teigen, 1988; Budescu & Wallsten, 1995; Dhami & Wallsten, 2005; Reagan et al., 1989)—and yet target propositions that had previously been presented with these qualifiers were rated equivalently on truth.

Studies 2a and 2b

One explanation for the results of Study 1 is that participants simply failed to remember the qualifiers two days later. Studies 2a and 2b were designed to shed light on this possibility. In both studies, only negative (e.g., *unlikely*, *impossible*) epistemic qualifiers were used. So, *all* qualifiers served to cast doubt the veracity of statements seen in the initial exposure phase, thereby reducing the recollective burden on participants.

Study 2a indirectly examined the role of memory for the qualifiers, considering that participants who remembered that all qualifiers were negative should be less susceptible to the illusory truth effect. Study 2b more directly examined memory for the qualifiers, with participants taking a short memory test for the qualifiers two days after initial exposure (as opposed to making truth judgments). Because of the similarity between the two studies, we report their materials and methods together, but

[1] Note that all 95% CIs reported in this manuscript are for the difference between means.

we then separately report the results of Studies 2a (truth ratings) and 2b (memory for the negative qualifiers).

Method

Participants. In Study 2a, 123 individuals ($M_{age} = 39.20$, $SD = 12.33$, range: 19–76, 70 females) voluntarily participated study via AMT and completed both sessions (attrition rate = 17%). In Study 2b, 42 individuals voluntarily participated via AMT and completed both sessions (attrition rate = 25%); one participant from Study 2b was excluded for failing to follow directions, so data were analyzed with the remaining 41 individuals ($M_{age} = 36.10$, $SD = 13.73$, range: 19–77, 17 females). In both studies, participant recruitment was restricted to individuals in the United States with a prior approval rating above 80%, and all participants reported being fluent English speakers. Participants from Study 1 were prevented from participating in Studies 2a and 2b. The sample size for Study 2a was determined by using a power analysis for equivalence testing of paired samples, as advocated by Chow et al. (2007).

Materials. The same 160 statements from Study 1 were used in both Studies 2a and 2b. However, all statements in both Studies 2a and 2b were qualified negatively during the initial exposure session, with *uncertain*, *unlikely*, *improbable*, or *impossible*.

Procedure. The procedure in Study 2a was identical to the procedure in Study 1, save for the use of solely negative qualifiers during exposure. The instructions and procedure in Study 2b differed from Study 2a only in the task completed two days after the initial exposure phase. In Study 2b, participants answered two items to probe their memory for the qualifiers from the first session, rather than rating the truth of old and new statements. Participants saw eight qualifiers on the computer screen at the same time (*certain*, *uncertain*, *likely*, *unlikely*, *probable*, *improbable*, *possible*, and *impossible)* and were asked to indicate which qualifiers had accompanied the statements in the first session. Participants rated confidence in their responses on a scale from 1 (*not at all confident*) to 5 (*very confident*). Then, participants selected the option that best described the qualifiers presented in session one: (1) all positive, (2) all negative, or (3) a mix of positive and negative. Participants rated their confidence in this categorization, using the same scale ranging from 1 (*not at all confident*) to 5 (*very confident*).

Results and Discussion

Study 2a: Truth ratings. The same equivalence testing procedures used in Study 1 were also used in Study 2a. The alpha level for all statistical tests was set at .05. Eighty-eight percent of participants reported not knowing whether any of the statements were true or false in the initial exposure phase based on prior knowledge. Our pattern of results did not differ after excluding participants who reported knowing whether at least one of the statements was true or false.

As in Study 1, we first investigated whether old statements—regardless of how those statements were qualified—were given higher truth ratings than new statements. A paired t test revealed that the average rated truth of old, previously qualified statements ($M = 4.88$, $SD = .75$) was significantly higher than that of new statements ($M = 4.65$, $SD = .71$), $t(122) = 6.66$, $p < .001$, 95% CI [.165, .305]. New statements first presented in the truth rating phase were not statistically equivalent in rated truth to old, previously qualified statements ($df = 123$; $t_1 = -9.98$, $p_1 < .001$; $t_2 = -3.33$, $p_2 > .999$). Because only negative epistemic qualifiers were used in the exposure phase, recalling a statement as old or familiar should be a reliable cue for doubting the veracity of the claim. These results suggest, however, that the negative qualifiers do not influence later judgments of truth.

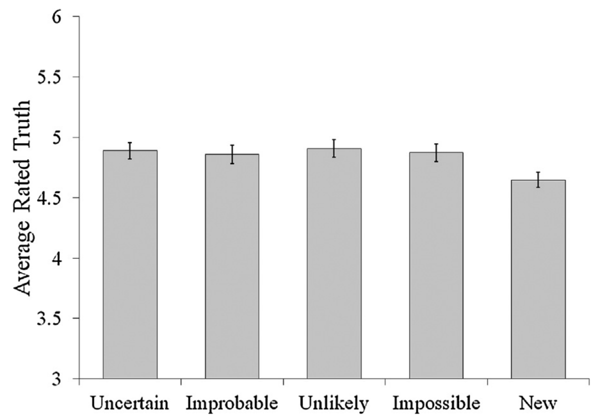

Figure 2. For Study 2, means and standard errors are depicted for the rated truth of both new statements and old statements grouped by qualifier originally seen accompanying the statement.

Next, we investigated whether truth ratings were statistically equivalent regardless of which qualifier accompanied the statements in the exposure phase (see Figure 2). To preview, the rated truth of old statements was statistically equivalent regardless of how they were qualified. Specifically, statements originally qualified with *uncertain* were statistically equivalent in rated truth to statements originally qualified with *unlikely* ($df = 122$; $t_1 = -3.70$, $p_1 < .001$; $t_2 = 2.96$, $p_2 = .002$), *improbable* ($df = 122$; $t_1 = -2.68$, $p_1 = .004$; $t_2 = 3.98$, $p_2 < .001$), and *impossible* ($df = 112$; $t_1 = -2.97$, $p_1 = .002$; $t_2 = 3.68$, $p_2 < .001$). Additionally, statements originally qualified with *unlikely* were statistically equivalent in rated truth to statements originally qualified with *improbable* ($df = 122$; $t_1 = -2.32$, $p_1 = .011$; $t_2 = 4.33$, $p_2 < .001$) and *impossible* ($df = 122$; $t_1 = -2.61$, $p_1 = .005$; $t_2 = 4.05$, $p_2 < .001$). Finally, statements originally qualified with *improbable* were statistically equivalent in rated truth to statements originally qualified with *impossible* ($df = 122$; $t_1 = -3.63$, $p_1 < .001$; $t_2 = 3.02$, $p_2 = .002$).

A subsequent ANOVA was computed with a single within-subjects factor (qualifier during the exposure phase: *uncertain*, *unlikely*, *improbable*, or *impossible*). The dependent variable was the rated truth of statements provided during the truth rating phase. There was no significant effect of qualifier on the rated truth of statements, $F(3, 366) = .39$, $p = .76$, partial $\eta^2 = .003$ (see Figure 2). Subsequent tests of pair-wise comparisons revealed no significant differences in rated truth as a function of the particular qualifier accompanying the statements during the exposure phase before correction for multiple comparisons (all *ps* > .31).

Overall, the pattern of results in Study 2a closely mirrored that from Study 1. The rated truth of statements was statistically equivalent regardless of the how the statements were originally qualified. And repeated statements were given *higher* truth ratings than new statements—despite all statements being negatively qualified in the initial exposure phase. In Begg and Armour (1991), each participant saw both negatively and positively biased statements in a single session. That is, participants did not exclusively see negatively biased statements along with new statements (i.e., no positively biased statements). In their studies, participants may have confused which statements were negatively biased and which were positively biased when judging truth. Mistaking previously negatively biased statements as having been positively biased could explain, at least partly, why previously negatively biased statements were rated as more likely to be true than new statements. In contrast, our Study 2a only used negatively qualified statements. Because no statements were positively qualified, participants could have just used a simple heuristic to judge truth: because all previously seen statements were negatively qualified, familiar or recognized statements are less likely to be true than new statements. Surprisingly, however, participants rated old statements, all of which had been negatively qualified, as more likely to be true than new statements.

Study 2b: Memory for negative qualifiers. For the first part of the memory test, participants indicated which of eight possible qualifiers they had seen in the initial exposure phase: 29.3% of participants correctly identified the four qualifiers (and only those four qualifiers) seen in the initial exposure phase. Participants who identified the four correct qualifiers were very confident in their judgments ($M = 4.25$, $SD = .75$). For the second part of the memory test, 51.2% of participants indicated that they had only seen negative qualifiers in the initial exposure phase; these participants were also very confident in their judgments ($M = 4.05$, $SD = .97$). 43.9% of participants indicated that they had seen a mix of positive and negative qualifiers in the initial exposure phase (confidence: $M = 3.83$, $SD = .79$). The remaining 4.9% of participants indicated that they had only seen positive qualifiers in the initial exposure phase.

Collectively, these results suggest that more than half of participants remembered that they had only seen negative qualifiers in the initial exposure phase. However, while performance was better than chance, a large contingent of participants thought that they had seen a mixture of positive and negative qualifiers in the initial exposure phase. Presumably, then, many participants in Study 2a could not use qualifying information to make truth judgments because they did not adequately encode the qualifying information in the initial exposure phase. Note that participants in Study 2b who completed the memory test for the qualifiers they saw in the first session did *not* also make truth judgements for the target propositions. This was a design choice to avoid possible contamination across the memory test and truth judgments. Thus, it is not possible to consider the truth ratings of participants who also remembered the qualifiers to a criterion. Studies 3a and 3b aimed to further boost memory for the qualifiers, to ensure that the results were not dependent on forgetting the qualifiers.

Studies 3a and 3b

Studies 3a and 3b required participants to demonstrate they had processed each qualifier. Immediately after reading each statement, they identified (from memory) which of the four qualifiers they had just seen: *uncertain*, *unlikely*, *improbable*, or *impossible*. Paralleling Studies 2a and 2b, Study 3a examined the effect of qualifying language on later truth judgements and Study 3b examined explicit memory for the qualifiers. Because of their similarities, the materials and methods of 3a and 3b are reported together, followed by the results for truth ratings (Study 3a) and qualifier memory (Study 3b), separately.

Method

Participants. In Study 3a, 148 individuals ($M_{age} = 35.05$, $SD = 11.14$, range: 18–71, 68 females) voluntarily participated in this study via AMT and completed both sessions (attrition rate = 12%). In Study 3b, 40 individuals voluntarily participated in this study via AMT and completed both sessions (attrition rate = 13%); two participants were excluded in Study 3b for failing to follow directions, so data were analyzed with the remaining 38 individuals ($M_{age} = 29.50$, $SD = 6.76$, range: 19–48, 15 females). In both studies, participant recruitment was restricted to individuals in the United States with a prior approval rating above 80%. All participants reported being fluent English speakers. Participants from previous studies were automatically prevented from participating. As before, the sample size was determined by using a power analysis for equivalence testing of paired samples, as advocated by Chow et al. (2007).

Materials. The materials in Studies 3a and 3b were the same as in Studies 2a and 2b: 160 statements, all of which were qualified negatively during the initial exposure session, with *uncertain*, *unlikely*, *improbable*, or *impossible*.

Procedure. The procedures were identical to those of Studies 2a and 2b, respectively, with one exception: immediately after rating interest in each statement, participants were required to identify (from memory) which of the four qualifiers they had just seen. They did so by selecting one of four options on the screen: *uncertain*, *unlikely*, *improbable*, or *impossible*. As before, in the second session, participants in Study 3a rated old and new statements for truth, and participants in Study 3b were tested on their memory for the qualifiers.

Results and Discussion

Studies 3a and 3b: Encoding attention check. On average, participants were excellent at identifying the just-seen qualifier. In Study 3a, they correctly identified the qualifier that they had just seen on 92.8% of trials; in Study 3b, they correctly identified the qualifier on 93.2% of trials. These results confirm that participants were processing qualifying information in the initial exposure phase.

Study 3a: Truth ratings. The same equivalence testing procedures used in Study 1 and Study 2a were also used in Study

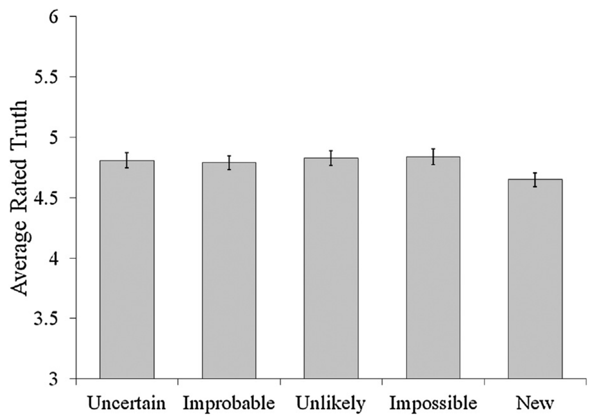

Figure 3. For Study 3a, means and standard errors are depicted for the rated truth of both new statements and old statements grouped by qualifier originally seen accompanying the statement.

3a. The alpha level for all statistical tests was set at .05. Eighty-three percent of participants reported not knowing whether any of the statements were true or false in the initial exposure phase based on prior knowledge. Our pattern of results did not differ after excluding participants who reported knowing whether at least one of the statements was true or false. Moreover, even after removing the subset of participants with the poorest performance on the encoding task (2.5 *SD*s below the mean; three participants), our pattern of results did not differ.

As in our previous studies, we first investigated whether old statements—regardless of how those statements were qualified—were given higher truth ratings than new statements. A paired *t* test revealed that the average rated truth of old, previously qualified statements ($M = 4.82$, $SD = .68$) was significantly higher than that of new statements ($M = 4.65$, $SD = .68$), $t(147) = 5.71$, $p < .001$, 95% CI [.11, .23]. New statements first presented in the truth rating phase were not statistically equivalent in rated truth to old, previously qualified statements ($df = 147$; $t_1 = -9.38$, $p_1 < .001$; $t_2 = -2.08$, $p_2 > .981$). Because only negative epistemic qualifiers were used in the exposure phase, recalling a statement as old or familiar should be a reliable cue for doubting the veracity of the claim. These results suggest, however, that the negative qualifiers do not influence later judgments of truth.

Next, we investigated whether truth ratings were statistically equivalent regardless of which qualifier accompanied the statements in the exposure phase (see Figure 3). As in Study 1 and Study 2a, truth ratings were statistically equivalent regardless of how those statements were originally qualified. Specifically, statements originally qualified with *uncertain* were statistically equivalent in rated truth to statements originally qualified with *unlikely* ($df = 147$; $t_1 = -4.10$, $p_1 < .001$; $t_2 = 3.20$, $p_2 = .001$), *improbable* ($df = 147$; $t_1 = -3.18$, $p_1 < .001$; $t_2 = 4.12$, $p_2 < .001$), and *impossible* ($df = 147$; $t_1 = -4.36$, $p_1 < .001$; $t_2 = 2.94$, $p_2 = .002$). Additionally, statements originally qualified with *unlikely* were statistically equivalent in rated truth to statements originally qualified with *improbable* ($df = 147$; $t_1 = -2.66$, $p_1 = .004$; $t_2 = 4.64$, $p_2 < .001$) and *impossible* ($df = 147$; $t_1 = -3.96$, $p_1 < .001$; $t_2 = 3.33$, $p_2 < .001$). Finally, statements originally qualified with *improbable* were statistically equivalent in rated truth to statements originally qualified with *impossible* ($df = 147$; $t_1 = -4.89$, $p_1 < .001$; $t_2 = 2.41$, $p_2 = .009$).

A subsequent ANOVA was computed with a single within-subjects factor (qualifier during the exposure phase: *uncertain*, *unlikely*, *improbable*, or *impossible*). The dependent variable was the rated truth of statements provided during the truth rating phase. There was no significant effect of qualifier on the rated truth of statements, $F(3, 441) = .57$, $p = .64$, partial $\eta^2 = .004$ (see Figure 3). Subsequent tests of pair-wise comparisons revealed no significant differences in rated truth as a function of the particular qualifier accompanying the statements during the exposure phase before correction for multiple comparisons (all ps $> .21$).

Overall, the pattern of results in Study 3a closely mirrored those from the previous studies, even though participants' responses during the encoding phase showed they had successfully processed the qualifying information. The rated truth of statements was statistically equivalent regardless of the qualifier originally accompanying the statement. And repeated statements were given *higher* truth ratings than new statements—despite all statements being negatively qualified in the initial exposure phase.

Study 3b: Memory for qualifiers. For the first part of the memory test, participants indicated which of eight possible qualifiers they had seen in the initial exposure phase: 71.1% of participants correctly identified the four qualifiers (and only those four qualifiers) seen in the initial exposure phase. Participants who identified the four correct qualifiers were very confident in their judgments ($M = 4.71$, $SD = .53$). For the second part of the memory test, 79.0% of participants indicated that they had only seen negative qualifiers in the initial exposure phase; these participants were also very confident in their judgments ($M = 4.67$, $SD = .71$). The remaining participants indicated that they had seen both positive and negative qualifiers in the first session; no participants believed that all qualifiers had been positive.

Collectively, these results suggest that most participants do remember that they only saw negative qualifiers in the initial exposure phase when probed—and yet the results from Study 3a suggest that this knowledge is not sufficient to eliminate the illusory truth effect.

Studies 4a and 4b

In the final pair of studies, we eliminate two possible alternative interpretations for our pattern of results. First, we wanted to eliminate any concern that participants might have been interpreting the qualifiers in some way other than what we intended. Second, we wanted to ensure that the qualifiers were associated with the target propositions in the exposure phase. Our prior modifications revealed that participants processed the qualifiers, but they did not show that they also processed the target proposition paired with each qualifier. Two changes to the procedure in the initial exposure phase were made in Studies 4a and 4b to eliminate these possible explanations. Moreover, we also only included the three most extreme negative qualifiers (i.e., unlikely, improbable, and impossible) from the previous studies to further ease the recollective burden for the qualifying information.

First, to ensure that all participants were interpreting the qualifiers as we intended, we revised the instructions so that they explicitly defined what the qualifiers signified. More specifically, we explained that words like *impossible*, *improbable* and *unlikely* are called qualifiers, and that these words are included toward the beginning of each trivia statement. Then, we explained that the purpose of these qualifiers is to cast doubt on the truth of the statements that they accompany. We also provided an example of a correct interpretation of the qualifying information: "If you see the statement, 'It is unlikely that the capital of Tuvalu is called Savu,' then you should suspect that the capital of Tuvalu is not called Savu." Participants were tested on these instructions to confirm that they understood how we wanted them to interpret the qualifiers.

Second, to ensure each qualifier was processed in relation to the target statement, we added a component to the attention check. During encoding, immediately after the presentation of each statement, participants were tested on their memory for both the qualifier (as in Studies 3a and 3b) and a content word from the target proposition. For example, immediately after reading the claim "It's unlikely that the body of a rotten tree is called a daddock" participants had to identify which qualifier had just been presented (out of 3 possible options: *unlikely*, *improbable*, and *impossible*) and which of two content words had been read (daddock or cambium).

Thus, Studies 4a and 4b were designed to replicate Studies 3a and 3b, all while eliminating concerns about how participants interpreted the qualifiers and ensuring that the key content was encoded in the context of specific qualifiers.

Method

Participants. A total of 144 individuals voluntarily participated in Study 4a via AMT and completed both sessions (attrition rate = 18%). Three participants were excluded for failing to follow instructions by using outside resources on the task, and three additional participants were excluded for failing to correctly answer the test question about the instructions. So, data for Study 4a were analyzed with the remaining 138 individuals (M_{age} = 40.84, *SD* = 12.54, range: 19–72, 67 females).

Sixty-two individuals voluntarily participated in Study 4b via AMT and completed both sessions (attrition rate = 18%). Two participants were excluded for skipping at least one question, and three additional participants were excluded for failing to correctly answer the test question about the instructions, so Study 4b data were analyzed with the remaining 57 individuals (M_{age} = 35.19, *SD* = 10.30, range: 20–67, 28 females).

In both studies, participant recruitment was restricted to individuals in the United States with a prior approval rating above 80%. All participants reported being fluent English speakers, and participants from the previous studies were prevented from participating.

Materials. A randomly selected subset of 120 items used in the previous studies were used in Studies 4a and 4b. All statements were qualified negatively during the initial exposure session, with *unlikely*, *improbable*, or *impossible*.

Procedure. After providing informed consent, participants in both Studies 4a and 4b were given more detailed instructions than in the previous studies. These instructions extensively described how they should interpret the qualifiers. Specifically, participants were told that words like *impossible*, *improbable*, and *unlikely* are called qualifiers, and that these words would be used in our study to cast doubt on the information they accompany; this same language was used in Begg and Armour (1991). Participants then read the following example: "If you see the statement, 'It is *unlikely* that the capital of Tuvalu is called Savu,' then you should suspect that the capital of Tuvalu is not called Savu." On the following page, participants were tested on these instructions by answering the following True/False question: "The purpose of qualifiers like *impossible*, *improbable*, and *unlikely* is to cast doubt on the truth of the statements that they accompany." Participants who responded False were excluded from the analyses, as described above.

Participants in both Studies 4a and 4b then completed the initial *exposure phase*, during which they rated 60 statements (each of the three qualifiers accompanied 20 different statements) for subjective interest on an 8-point scale ranging from 1 (*very interesting*) to 8 (*very uninteresting*). Response submission for each statement was only available after 3 s had passed to encourage processing. Critically, immediately after rating interest in each statement, participants were required to identify (from memory) on the following page which of the three qualifiers they had just seen. They did so by selecting one of three options on the screen: *unlikely*, *improbable*, or *impossible*. On that same page, participants were also required to identify (from memory) a key word they had just seen in the target proposition on the previous page by selecting one of two possible options. The included lure word was carefully matched with the target word based on extensive norming (Wang et al., 2016). We counterbalanced across subjects which of these two words served as the target or the lure.

Two days after the initial exposure phase, participants in Study 4a completed the *truth rating phase*, and participants in Study 4b completed the *memory test.* The final sessions differed from the earlier experiments only in the number of items, given that only the three most extreme negative qualifiers were used in Studies 4a and 4b. Thus, the truth test involved rating 120 items (60 old and 60 new) and the memory test required participants to select which qualifiers (out of six) that they had seen during the encoding phase: *likely*, *unlikely*, *probable*, *improbable*, *possible*, and *impossible*).

Results and Discussion

The alpha level for all statistical tests was set at .05. Eighty-two percent of participants reported not knowing whether any of the statements were true or false in the initial exposure phase based on prior knowledge. Our pattern of results did not differ after excluding participants who reported knowing whether at least one of the statements was true or false.

Studies 4a and 4b: Understanding of instructions. As noted earlier, 97.9% of participants in Study 4a and 95.2% of participants in Study 4b demonstrated understanding of the instructions about how to interpret the qualifiers. Only these participants were included in the analyses that follow.

Studies 4a and 4b: Encoding attention check. Critically, participants in both studies demonstrated successful encoding of the qualifiers in the initial exposure phase. On average, participants in Study 4a correctly identified the qualifier they had just seen on 95.2% of trials; performance was similarly high in Study 4b (94.4% correct). Moreover, participants demonstrated successful encoding of the content of each target proposition in the initial exposure phase, correctly identifying almost all of them (Study 4a: 98.0%; Study 4b: 97.3%). Thus, in both studies participants demonstrated encoding of both the qualifiers and the content of the target propositions.

Study 4a: Truth ratings. As in the previous studies, we first investigated whether old statements—regardless of how those statements were qualified—were given higher truth ratings than new statements. A paired *t* test revealed that the average rated truth of old, previously qualified statements ($M = 4.78$, $SD = 1.00$) was significantly higher than that of new statements ($M = 4.56$, $SD = .82$), $t(137) = 3.86$, $p < .001$, 95% CI [.11, .33]. New statements first presented in the truth rating phase were not statistically equivalent in rated truth to old, previously qualified statements ($df = 137$; $t_1 = .346$, $p_1 > .635$; $t_2 = 7.39$, $p_2 < .001$). Because only negative epistemic qualifiers were used in the exposure phase, recalling a statement as old or familiar should be a reliable cue for doubting the veracity of the claim. These results suggest, however, that the negative qualifiers do not influence later judgments of truth. Moreover, as in Study 3a, even after removing the subset of participants with the poorest performance on the encoding task (2.5 *SD*s below the mean; three participants), our pattern of results did not differ.

Next, we investigated whether truth ratings were statistically equivalent regardless of which qualifier accompanied the statements in the exposure phase (see Figure 4), using the same equivalence testing procedures that were used in the previous studies. As before, the rated truth of old statements was statistically equivalent regardless of how they were qualified. Specifically, statements originally qualified with *unlikely* were statistically equivalent in rated truth to statements originally qualified with *improbable* ($df = 137$; $t_1 = -3.39$, $p_1 < .001$; $t_2 = 3.65$, $p_2 < .001$) and *impossible* ($df = 137$; $t_1 = -3.96$, $p_1 < .001$; $t_2 = 3.09$, $p_2 = .001$). Additionally, statements originally qualified with *improbable* were statistically equivalent in rated truth to statements originally qualified with *impossible* ($df = 137$; $t_1 = -4.16$, $p_1 < .001$; $t_2 = 2.88$, $p_2 = .002$).

A subsequent ANOVA was computed with a single within-subjects factor (qualifier during the exposure phase: *unlikely*, *improbable*, or *impossible*). The dependent variable was the rated truth of statements provided during the truth rating phase. There was no significant effect of qualifier on the rated truth of statements, $F(2, 274) = .186$, $p = .830$, partial $\eta^2 = .001$ (see Figure 4). Subsequent tests of pair-wise comparisons revealed no significant differences in rated truth as a function of the particular qualifier accompanying the statements during the exposure phase before correction for multiple comparisons (all $ps > .51$).

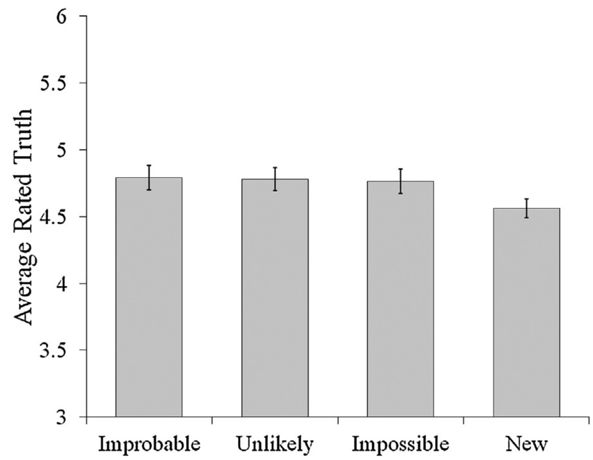

Figure 4. For Study 4a, means and standard errors are depicted for the rated truth of both new statements and old statements grouped by qualifier originally seen accompanying the statement.

Study 4b: Memory for qualifiers. On the first part of the memory test, participants indicated which of six possible qualifiers they had seen in the initial exposure phase: 78.9% of participants correctly identified the three qualifiers (and only those three qualifiers) seen in the initial exposure phase. Participants who identified the three correct qualifiers tended to be very confident in their judgments ($M = 4.73$, $SD = .58$). For the second part of the memory test, 80.7% of participants indicated that they had only seen negative qualifiers in the initial exposure phase; these participants also tended to be very confident in their judgments ($M = 4.76$, $SD = .57$). Of the remaining participants, ten indicated that they had seen both positive and negative qualifiers in the first session; one participant believed that all qualifiers had been positive.

Like Study 3b, these results from Study 4b suggest that most participants remembered which specific qualifiers they had seen in the initial exposure phase when probed—and yet the results from Study 4a suggest that this knowledge is not sufficient to eliminate the illusory truth effect.

General Discussion

We investigated whether truth judgments reflect how statements were originally qualified. Our results suggest that epistemic qualifiers do not influence truth judgments two days later. Moreover, the truth of previously qualified statements was rated equivalently, even when the qualifiers accompanying the statements had distinctly opposite meanings (i.e., *certain* and *impossible*). This finding is particularly noteworthy given that interpretations of extreme anchoring terms like *certain* and *impossible* tend to be highly consistent and reliable within and between individuals (Beth-Marom, 1982; Brun & Teigen, 1988; Budescu & Wallsten, 1995; Dhami & Wallsten, 2005; Reagan et al., 1989). Upon comprehension, propositions qualified with *certain* should be judged to be true, and propositions qualified with *impossible* should be judged to be false. Strikingly, in Studies 2a, 3a, and 4a, qualifiers did not differentially influence truth ratings, and old, previously qualified propositions were judged as more likely to be true than new propositions, even though *all* statements in these studies had been qualified to cast doubt on their veracity (e.g., *unlikely*, *improbable*, or *impossible*).

Forgetting of the qualifying information from the initial exposure phase is unlikely to drive our effects. Studies 3a and 4a explicitly tested participants' memory for the qualifiers immediately after reading each statement, to ensure that the qualifiers were encoded. Two days later, across Studies 2b, 3b, and 4b, 51–81% of participants were aware that all previously seen qualifiers had been negative. Participants were significantly more likely to remember that all qualifiers were negative (as opposed to only positive or a mix of positive and negative) in Studies 3b and 4b than they were in Study 2b ($ps < .05$). But the average differences in truth judgments between repeated (and previously qualified) and non-repeated statements were very similar between Studies 2a, 3a, and 4a (see 95% CIs between studies). That is, despite procedural changes to ensure encoding of and memory for the qualified statements, the illusory truth effect persisted at a similar magnitude across all studies. In other words, most people had the relevant knowledge about qualifiers stored in memory, and yet they failed to use it when making truth judgments (a form of *knowledge neglect*; see Umanath and Marsh, 2014).

These results are surprising given that prior work has shown that people do rely on source information when judging truth (Begg et al., 1992; Brown & Nix, 1996; Unkelbach & Stahl, 2009). The illusory truth effect is less pronounced when people remember that a low credibility source (as opposed to a higher credibility source) was associated with the original claim (Mitchell et al., 2005). One might expect epistemic qualifiers to function similarly, with the prediction that epistemic qualifiers will influence truth ratings until they are forgotten. However, our results from Studies 3b and 4b do not support this intuition. Our results are more consistent with Begg and Armour (1991), who found a larger illusory truth effect for positively biased statements than negative ones after a few minutes—but they still found an illusory truth effect for negatively biased statements. That is, memory for a statement's bias was insufficient to fully explain the overall pattern of results. We observed a more extreme pattern in our own data, whereby two days later the qualifiers had no impact on truth ratings. This more extreme pattern of results is more consistent with a fluency interpretation: instead of relying on qualifying information to make truth judgments, prior exposure made the target propositions easier to read in the truth rating phase, which in turn likely drove the truth judgments. Although these effects in our studies were not large, their size is consistent with prior research on the illusory truth effect (Dechêne et al., 2010).

Moreover, our pattern of results accords with recent findings suggesting that certain strategies for correcting common myths and misperceptions fail for systematic reasons. Asserting that a particular myth or misperception is false often backfires, as repeating the assertion increases fluency, reinforcing the truth of the claim (Schwarz, Newman, & Leach, 2016). These backfire effects may become more pronounced after a few days have passed (Peter & Koch, 2016). Individuals even believe that contradictory statements with the same surface-level appearance as others presented a week earlier are more likely to be true than completely novel statements (Garcia-Marques, Silva, Reber, & Unkelbach, 2015). Similarly, in our studies repeating statements likely increased fluency for the content of the propositions—despite the information about the veracity of the statements contained in the qualifiers—making fluency the primary driver of truth judgments for unqualified claims later.

Our findings suggest that attempting to rigorously and precisely define qualifiers will be insufficient to prevent errors in communication. Considerable research has highlighted the degree of variability in the interpretation of certain epistemic qualifiers (Brun & Teigen, 1988; Budescu & Wallsten, 1995; Dhami & Wallsten, 2005) and investigated ways to diminish errors and misunderstandings in communicating with qualifiers. For example, one approach involves standardizing the language (Hamm, 1991; Kadane, 1990; Mosteller & Youtz, 1990), whereas another approach seeks to establish conversion methods for relating different people's lexicons (Karelitz & Budescu, 2004). Unfortunately, even if it is possible to precisely define and standardize the meanings of epistemic qualifiers (and to minimize inter-individual variability in the interpretation of these qualifiers), our work suggests that serious failures in communication will persist. Qualifiers do not influence truth judgments two days later.

This work highlights a unique way in which false beliefs and misconceptions can enter one's knowledge base with potentially serious, negative consequences. Decision makers in high-stakes contexts—including those operating in medicine, law, finance, and intelligence—rely on epistemic qualifiers to communicate and comprehend critical information. If qualifying information no longer informs truth judgements after two days have passed, then individuals may not possess the requisite information to make informed, accurate decisions. These shortcomings in high-stakes decision-making could have serious negative ramifications for the well-being of the decision-maker and others. Future research will develop strategies for more deeply encoding, successfully recalling, and actively using qualifying information to make judgments about truth.

Conflict of Interest Statement

The authors declare no conflict of interest.

Author Contributions

All authors contributed to the study concept and design. Data collection and analysis were performed by M. L. Stanley. M. L. Stanley and B. W. Yang drafted the manuscript, and E. J.

Marsh provided critical revisions. All authors approved the final version of the manuscript for submission.

References

Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*(6), 576–605.

Begg, I., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*, 446–458.

Begg, I., & Armour, V. (1991). Repetition and the ring of truth: Biasing comments. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *23*, 195–213.

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *17*(3), 199.

Beth-Marom, R. (1982). How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*, 257–269.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, *20*(3), 285–293.

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088.

Brun, W., & Teigen, K. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior & Human Decision Processes*, *41*, 390–404.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, D. L. Medin, & R. Hastie (Eds.), *Decision making from a cognitive perspective* (pp. 275–318). New York: Academic Press.

Caton, C. E. (1966). On the general structure of the epistemic qualification of things said in English. *Foundations of Language*, *2*, 37–66.

Chow, S. C., Wang, H., & Shao, J. (2007). *Sample size calculations in clinical research*. CRC Press.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257.

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition*, *33*, 1057–1068.

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002.

Garcia-Marques, T., Silva, R. R., Reber, R., & Unkelbach, C. (2015). Hearing a statement now and believing the opposite later. *Journal of Experimental Social Psychology*, *56*, 126–129.

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes*, *48*, 193–223.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behavior*, *16*, 107–112.

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Lulu.com

Kadane, J. B. (1990). Comment: Codifying chance. *Statistical Science*, *5*, 18–20.

Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied*, *10*, 25–41.

Lakens, D. (2017). Equivalence tests: A practical primer for *t*-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*.

Mara, C. A., & Cribbie, R. A. (2012). Paired-samples tests of equivalence. *Communications in Statistics – Simulation and Computation*, *41*(10), 1928–1943.

Marslen-Wilson, W. D., & Tyler, L. K. (1997). Dissociating types of mental computation. *Nature*, *387*, 592–594.

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*(5), 424–428.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

Mitchell, J. P., Dodson, C. S., & Schacter, D. L. (2005). fMRI evidence for the role of recollection in suppressing misattribution errors: The illusory truth effect. *Journal of Cognitive Neuroscience*, *17*(5), 800–810.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science*, *5*, 2–16.

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, *13*(2), 225–253.

Peter, C., & Koch, T. (2016). When debunking scientific myths fails (and when it does not): The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication*, *38*(1), 3–25.

Reagan, T. R., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*, 433–442.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680.

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95.

Teigen, K. H., & Brun, W. (2000). Ambiguous probabilities: When does $p = 0.3$ reflect a possibility, and when does it express a doubt? *Journal of Behavioral Decision Making*, *13*, 345–362.

Umanath, S., & Marsh, E. J. (2014). Understanding how prior knowledge influences memory in older adults. *Perspectives on Psychological Science*, *9*, 408–426.

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230.

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*(1), 22–38.

Wallsten, R. S. (1990). Measuring vague uncertainties and understanding their use in decision making. In G. M. Von Furstenberg (Ed.), *Acting under uncertainty: Multidisciplinary conceptions* (pp. 377–398). London: Kluwer.

Wallsten, T. S., Budescu, D. V., & Tsao, C.-Y. (1997). Combining linguistic probabilities. *Psychologische Beitraege*, *39*, 27–55.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, *31*(2), 135–138.

Wang, W. C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. *Journal of Cognitive Neuroscience*, *28*, 739–746.

West, R. F., & Stanovich, K. E. (1988). How much of sentence priming is word priming? *Bulletin of the Psychonomic Society*, *26*, 1–4.

Received 27 April 2018;
received in revised form 5 July 2018;
accepted 13 August 2018
Available online 13 September 2018